Rule Extraction from Support Vector Machines: Measuring the Explanation Capability Using the Area under the ROC Curve

Nahla Barakat[±*] and Andrew P. Bradley[*]
[±]*Sohar University, Oman*
[*]*School of Information Technology and Electrical Engineering (ITEE)*
The University of Queensland, St Lucia, QLD 4072, Australia
{nahla|bradley}@itee.uq.edu.au

Abstract

Recently, the area of rule extraction from support vector machines (SVMs) has been explored. One important indication of the success of a rule extraction method is the performance of extracted rules as compared to the original SVM. In this paper, we describe the use of the area under the receiver operating characteristics (ROC) curve (AUC) to assess the quality of rules extracted from an SVM. In particular, we directly compare AUC to the more commonly used measures of accuracy and fidelity and show that AUC is both a more reliable and meaningful measure to use.

1. Introduction

Support vector machines (SVMs) have shown superior performance in a variety of application areas including medical diagnosis. However, one limitation of SVMs is that they produce *black-box* models with no real explanation of the classification decisions being made. This is problematic, especially in medical applications, where diagnostic decisions may have very real ethical and clinical implications. Therefore, there has been a variety of methods proposed to extract rules from a trained SVM e.g., [1], [2], [3], [4]. The primary goal of this work has been to construct rule based classifiers that can explain the classifications made by an SVM.

The task of rule extraction is one of expressing the knowledge acquired by the SVM during the training process in a comprehensible form easily understood by end-users. Fung *et al*. argue that even limited explanation power can provide a valuable check of the internal logic of a *black-box* model and can positively influence the acceptance of these models in daily clinical practice [4]. One of the most important indications on the success of a rule extraction method is the quality of the extracted rules, and the extent to which they have a same performance as the SVM from which they were extracted. Put another way, the rules can only provide an explanation for the SVM if their performance is, in some sense, equivalent. Therefore, the problem here is to compare the performance of two different classifiers: the SVM and the extracted rule set; and to test for equivalence in measured performance. Clearly, a complementary measure of rule set quality is their *interpretability* by domain experts. However, in this paper we focus purely on the prerequisite to this type of analysis.

The most commonly used measures for rule quality are accuracy and fidelity [1], [2], [3]. Accuracy is the percentage of a correct classification, whilst fidelity measures the extent to which the prediction behavior of a rule set mimics that of the SVM. It has been shown that accuracy is not a reliable criteria for comparing the performance of two classifiers as it does not cater for skewed class priors or unequal misclassification costs [5], [6]. Clearly, these arguments also apply to fidelity as it is a direct pair-wise comparison of the accuracy of the SVM and the extracted rule set. In addition, even performance measures, such as the true positive rate (TPR) and false positive rate (FPR), which are not affected by class priors, can also be difficult to interpret and compare. For example, when a rule set is extracted from an SVM it is highly likely that the two classifiers will operate at different TPRs *and* FPRs, despite the fact that both classifiers ideally represent the same learned knowledge. Comparing two classifiers that operate at different TPRs and FPRs is complicated by the fact that it is not clear from these single points if the differences in TPR and FPR reflect a true difference in classification performance or equivalent performance, just at different operating points. That is, unless an increased TPR is associated with a decreased FPR then these classifiers could be different points on the same receiver operating characteristic (ROC) curve. Clearly, this is not the case if the two classifiers

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

operate, either at the same TPR or the same FPR, but as we will show in the result section, these situations are unlikely to happen in practice. Therefore, in this paper we propose to compare the performance of the SVM and its associated rule set over all operating points (and therefore all misclassification costs) using the area under the ROC curve (AUC).

The paper is organized as follows: first we provide a brief introduction to SVMs, followed by a definition of the ROC curve and AUC. Next, we summarize the eclectic approach to rule extraction from SVMs, outline the proposed experimental methodology and present a discussion of our results and conclusions.

2. Support Vector Machines

SVMs are based on the principle of structural risk minimization, which aims to minimize the true error rate. SVMs operate by finding a linear hyper-plane that separates the positive and negative examples with a maximum interclass distance or *margin*. In the case of non-separable data, a soft margin hyper-plane is defined to allow errors ξ_i (slack variable) in classification. Hence, the optimization problem is formulated as follows [7]:

$$\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i$$

$$\text{Subject to } y_i(wx_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0$$

Where C is a regularization parameter which defines the trade-off between the training error and the margin [7]. In the case of non-linearly separable data, SVMs map input data to be linearly separable in the feature space using kernel functions. Including kernel functions, and Lagrange multiplier α_i, the dual optimization problem is modified as follows [7]:

$$\text{maximize } w(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{l}\alpha_i y_i \alpha_j y_j K(x_i . x_j)$$

$$C \geq \alpha_i \geq 0 \quad \forall_i, \qquad \sum_{i=1}^{l}\alpha_i y_i = 0$$

In the case of unequal misclassification costs, a cost factor $J\ (C_+/C-)$ is introduced, by which training errors on positive examples outweigh errors on negative examples[11]. Therefore, optimization problem becomes:

$$\text{minimize } \frac{1}{2}\|w\|^2 + C_+\sum_{i:y_i=1}\xi_i + C_-\sum_{j:y_j=-1}\xi_j \qquad (5)$$

$$\text{Subject to } y_k(wx_k + b) \geq 1 - \xi_k, \qquad \xi_k \geq 0$$

3. The ROC Curve

In signal detection theory, the plot of TPR against FPR as the decision threshold is varied is known as a ROC curve. Therefore, a classifier with a single TPR and FPR corresponds to a single operating point on a ROC curve [8]. ROC curves describe the predictive behavior of a classifier independent of class distributions and error costs [6]. The area under the ROC curve (AUC) has been shown to be a useful and sensitive measure of classifier performance and varies between 0.5 (random guessing) and 1.0 (perfect classifier) [6], [8].

4. Eclectic Rule Extraction from SVMs

In this paper, we use the eclectic rule extraction approach described in [1]. This approach uses a labeled data set to train an SVM and get an SVM (classifier) with acceptable accuracy, precision, and recall. Next, a synthetic data set composed of the patterns that became support vectors is constructed with the target class for these patterns replaced by the class predicted by the SVM. Rules representing the concepts learned by the SVM are then extracted from this synthetic data set using the C5 decision tree learner [9].

5. Experimental Methodology

Experiments were performed using four benchmark medical data sets from [10] to assess the quality of the rules extracted using the eclectic rule-extraction approach. The details are as follows:

- **Pima Indians diabetes**: A sample of 438 patterns were used from the original data set, after removing all patterns with a zero value for the attributes 2-hour OGTT plasma glucose, diastolic blood pressure and triceps skin fold thickness which are clinically insignificant;
- **Heart diseases**: The reduced Cleveland heart diseases data set was used. All patterns with missing values were discarded;
- **Breast cancer**: The Wisconsin breast cancer data set was used. All repeated patterns were discarded to avoid the bias resulting from the boosting effect of those patterns;
- **Dermatology**: This database has six classes and so our experiments were conducted as a binary classification task of class 1 (psoriasis) against the other five classes. All patterns with missing values were discarded.

Table 1: Data Sets

Data set	No. Features	Training set size	Test set size
Pima Indians	8	247	191
Breast cancer	9	208	147
Heart Disease	13	223	74
Dermatology	34	173	193

As shown in Table 1, each of the four data sets was split into disjoint training and test sets. The rules were then extracted directly from each trained SVM and the

test set was then used to estimate the generalization performance of the SVM and extracted rule set in terms of accuracy, fidelity, TPR, FPR and AUC. SVM^{light} [12] was used in all experiments as follows:

1. Leave-one-out cross validation was used to select the training parameters (kernel type and the regularization parameter C) that minimized error rate over the training set;
2. A number of SVM models were generated by varying the misclassification cost factor, *J*, starting with small value and increasing *J* until no change in TPR or FPR was observed;
3. Each of the generated SVM models were then used to classify the independent test set and accuracy, TPR and FPR were calculated;
4. Rules were then extracted from each of the SVM models using the eclectic rule extraction [1];
5. Each of the extracted rule sets were then used to classify the same independent test set and again accuracy, TPR and FPR were computed;
6. ROC curves were then plotted for both the SVM and rule sets and AUC computed using trapezoidal integration. Standard errors for the AUC were estimated via the standard error of the Wilcoxon statistic [6].

It should be noted that in three of the four data sets the ROC curve had to be manually connected to the point (1,1). Although it is known that trapezoidal integration systematically underestimates AUC [6] we minimized this effect by ensuring that we had at least seven points (and normally up to 15) from which to estimate the ROC curve. Additionally, trapezoidal integration does not rely on any assumptions as to the underlying distributions of the positive and negative examples.

6. Results and Discussion

Accuracy and fidelity results comparing the SVM and the extracted rules are shown in Table 2. The standard deviation of accuracy being estimated using the Binomial distribution [13]. It can be seen that the highest fidelity is obtained on the Breast cancer and Dermatology data sets. This can be attributed to the fact that these data sets have discrete valued features and so map well to rule sets. Table 3 shows the TPR and FPR for equal misclassification costs (J = 1).

The ROC curves for these data sets are shown in Figures 1 to 4 whilst the AUC and associated standard errors are shown in Table 4. Comparing the ROC curves for the SVM and the extracted rule sets, it can be seen that both curves follow the same pattern with increasing J. However, as can be seen in Table 3, at same value of J, a rule set with a lower TPR than the SVM is only associated with an increase in FPR on two of the data sets (Pima Indians and Dermatology). Therefore, it is only by plotting the complete ROC curve that we can make an informed judgment as to true their relative performance.

Table 2: Accuracy and fidelity (J = 1).

Data set	SVM	Rules	Fidelity
	Acc. ± Std	Acc. ± Std	
Pima Indians	0.78 ± 0.03	0.88 ± 0.02	0.82
Breast cancer	0.94 ± 0.02	0.91 ± 0.02	0.97
Heart Disease	0.74 ± 0.05	0.82 ± 0.04	0.88
Dermatology	1.00 ± 0.00	0.98 ± 0.01	0.98

Table 3: True and false positive rates (J = 1).

Data set	SVM		Rules	
	TPR	FPR	TPR	FPR
Pima Indians	0.58	0.00	0.71	0.02
Breast cancer	0.90	0.03	0.86	0.01
Heart Disease	0.91	0.45	0.97	0.43
Dermatology	1.00	0.00	0.97	0.02

Table 4: The area under the ROC curve.

Data set	SVM AUC ± Std	Rules AUC ± Std
Pima Indians	0.82 ±0.03	0.94 ± 0.02
Breast cancer	0.97 ± 0.01	0.96 ± 0.02
Heart Disease	0.89 ± 0.04	0.81 ± 0.051
Dermatology	1.00 ± 0.00	0.984 ± 0.011

The AUC for the SVM is greater than that for the rule sets in three of the four data sets. The exception to this is Pima Indians where the test set is noisier and so the induction bias of C5, producing shorter tree with maximum information gain, is beneficial [9]. To determine if the differences in AUC between the SVM and the extracted rule sets are statistically significant a large sample z test was performed [13]. Results indicate that the null hypothesis, that there is no difference in measured AUC, can not be rejected ($p > 0.05$) on three out of the four data sets. Only on Pima Indians is there a significant difference in AUC, which indicates that the rule set AUC is greater than that of the original SVM ($p < 0.005$). Using accuracy as the sole performance measure shows a significant difference ($p < 0.05$) between the SVM and the rule sets on three of the four data sets; only on Breast cancer do the SVM and rule set have equivalent accuracy. On the Dermatology data set in particular, this difference is almost certainly misleading once the similarity of the ROC curves in Figure 4 has been observed. Although the ROC curve for Heart Disease, in Figure 3, is harder to interpret, AUC indicates that, over all operating points, these ROC curves are equivalent.

It is also worth noting that the rules extracted from the SVM are largely comprehensible, having a maximum of five rules and four antecedents.

7. Conclusions

In this paper we have described the use of AUC to assess the quality of rules extracted from an SVM. We have shown that ROC curves and AUC provide a more reliable measure for assessing the quality of the extracted rules than the commonly used measures of accuracy and fidelity. In addition, we have shown that on all four data sets the extracted rules have at least an equivalent performance to the original SVM.

References

[1] N. Barakat and J. Diederich. "Eclectic rule-extraction from support vector machines", *Int. Journal Computational Intelligence,* 2 (1), 2005, pp. 59-62.

[2] H. Núñez, C. Angulo, A. Catala. "Rule-extraction from support vector machines", *Proc. European Symposium on Artificial Neural Networks*, 2002, pp. 107-112.

[3] Y. Zhang, H. Su, T. Jia, J.Chu. "Rule extraction from trained support vector machines", *Proc. Advances in Knowledge Discovery and Data Mining: 9th Pacific-Asia Conference PAKDD*, Springer, 2005, pp. 61-70.

[4] G. Fung, S. Sandilya, R. Rao. "Rule extraction from linear support vector machines", *Proc. of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2005.*

[5] F. Provost, T. Fawcett, R. Kohavi, "The case against accuracy estimation for comparing induction algorithms", *Proc. International Conference on Machine Learning*, 1998, pp. 445-453.

[6] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recognition,* 30(7), 1997, pp 1145-1159.

[7] C. Burges. "A tutorial on support vector machines for pattern recognition", *Data Mining and Knowledge Discovery*, Kluwer Academic, 1998.

[8] T. Fawcett "ROC graphs: notes and practical considerations for researchers", Kluwer Academic Publishers, Netherlands, 2004.

[9] Data Mining Tools See5 and C5.0, Rule Quest data mining tools, http://www.rulequest.com

[10] C. Merz and P. Murphy, UCI machine learning repository, Irvine, 1992.
http://www.ics.uci.edu/~mlearn/MLRepository.html

[11] K. Morik, P. Brockhausen, T. Joachims, "Combining statistical learning with knowledge-based approach-A case study in intensive care monitoring", *Proc. European Conference on Machine Learning,* Springer, 1998.

[12] T. Joachims. "Making large-scale SVM learning practical," *Advances in kernel Methods-Support Vector Learning*, B.Schölkopf, C. Burges, and A. Smoland (Eds.), MIT-Press, 1999. http://svmlight.joachims.org

[13] A. P. Bradley and I.D. Longstaff, "Sample size estimation using the receiver operating characteristic curve", *Proc. 17th International Conference on Pattern Recognition*, Cambridge, UK, Vol. 4, 2004, pp. 428-431.

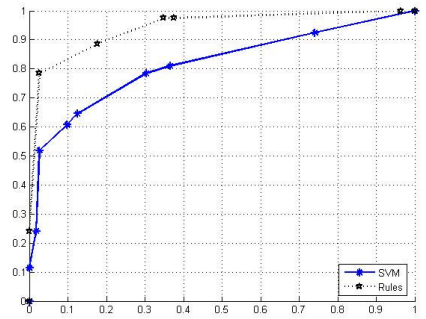

Figure 1: ROC curve for Pima Indians.

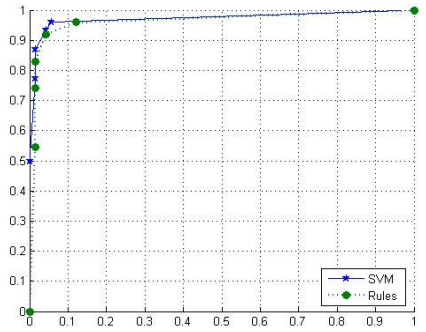

Figure 2: ROC curve for Breast Cancer.

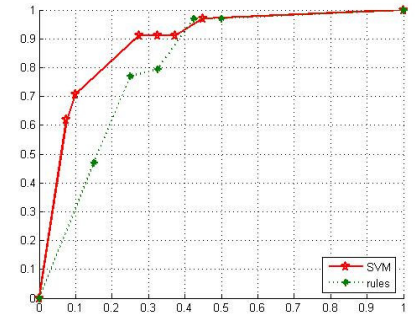

Figure 3: ROC curve for Heart Disease.

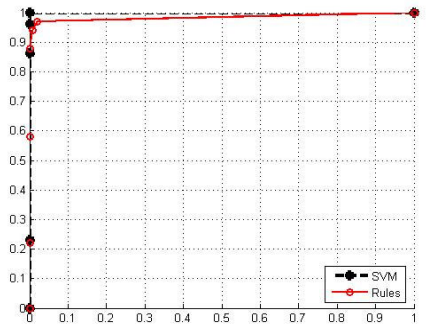

Figure 4: ROC for Dermatology.